Towards New Measures of Information Retrieval Evaluation

William R. Hersh, M.D.
Diane L. Elliot, M.D.
David H. Hickam, M.D., M.P.H.
Stephanie L. Wolf
Oregon Health Sciences University, Portland, OR

Anna Molnar
Christine Leichtenstien
University of Ulm, Ulm, Germany

All of the methods currently used to assess information retrieval (IR) systems have limitations in their ability to measure how well users are able to acquire information. We utilized a new approach to assessing information obtained, based on a short-answer test given to senior medical students. Students took the ten-question test and then searched one of two IR systems on the five questions for which they were least certain of their answer Our results showed that pre-searching scores on the test were low but that searching yielded a high proportion of answers with both systems. These methods are able to measure information obtained, and will be used in subsequent studies to assess differences among IR systems.

INTRODUCTION

As information retrieval (IR) systems proliferate, it is necessary to assess their usefulness to clinicians. The most common approach for evaluating IR systems has been to measure usage frequency and/or user satisfaction. While usage frequency is easy to measure, it provides no insight into why the system was used or how successful the user was in finding information. Likewise, user satisfaction does not elucidate how users interact with IR systems and usually has not indicated user perceptions of the system's cost-effectiveness. Indeed, it has been shown that over a third of clinician users stopped using Grateful Med during a several-year period [1], and that usage dropped by two-thirds when access fees were imposed [2].

The next level of retrieval evaluation has been to measure users' success at retrieving relevant documents using indices such as recall and precision. While these indices provide a starting point at determining how much useful information is obtained from an IR system, they inherently are based on judgments of the relevance of documents to users' queries. Yet relevance is difficult to measure. Not only is interobserver agreement in relevance judgments low [3, 4, 5], but judgments of relevance are influenced by factors such as document order and expertise of the judge [6, 7].

Even if relevance judgments were valid, there are other problems with using recall and precision as measures of information retrieval [8]. There is the practical question about the definition of a retrieved document. We have seen users in previous studies who started out with a poor search, retrieving a large number of nonrelevant documents, but later refining the search to retrieve many relevant documents. In some cases, the poor search was just due to a typing error. Yet despite the ultimate success of the search, the recall and precision values were poor.

In the case of recall, the user's retrieval of more documents does not necessarily correlate with better searching success, as there is a great deal of redundancy in the medical literature. A related problem arises in the context of assessing differences between IR systems. In particular, how well do recall and precision actually represent "performance" of an IR system? Should the proportion of relevant documents obtained in the collection (recall) or search (precision) be the "gold standard" for performance? Knowing the quantity of relevant articles tells us nothing of the quality. It fails to indicate whether the information need that prompted the search was satisfied. Furthermore, when comparing two systems, while we may be able to show statistical significance between the results (with a t-test or some other appropriate statistical measure), we have no idea what constitutes a clinically significant difference.

To explore the feasibility of an alternative method to evaluate how well IR systems help users meet their information needs, we utilized an alternative approach, adapting methods previously used to evaluate a hypertext statistical textbook [9], a historical encyclopedia [10], and a series of biomedical factual databases [11]. The goal of adopting this approach was to assess how well users answered clinical questions with an IR system. The purpose of this study was to determine whether this method could measure information acquisition and thus be used as a method to determine the effectiveness of user interaction with the system.

0195-4210/94/$5.00

METHODS

For this study we used two IR programs developed at Oregon Health Sciences University (OHSU). The first of these was SWORD, which features a natural language searching interface with relevance ranking. With SWORD, the user enters a free-text query and retrieved documents are ranked based on the "similarity" of their words to those in the query [12]. The second program was BOOLEAN, which utilizes a Boolean interface modeled after the Grateful Med system, where the words within each line are connected by logical OR, followed by the connection of each line with logical AND [5]. Both programs log every interaction with the user, including submitting a query, selecting a document to view, and browsing other documents. The database searched by both programs was an electronic version of the textbook, *Scientific American Medicine* [13], divided into over 6,600 "documents" based upon the hierarchical structure of the print version.

To measure information acquisition, we developed a ten-question short-answer test at the senior medical student level of difficulty (Table 1). The test questions were designed to have specific answers in the database, so that we had at least one document that provided the "answer" to each question. The test was given before and after searching, with the measurements of difference assessed by correctness of answers as well as changes in certainty of the answer.

All medical students from the senior class at Oregon Health Sciences University were sent a letter asking them to participate, of which 13 volunteered. Each student completed a brief questionnaire asking about prior computer experience, and we also obtained each student's class rank from the OHSU Dean's office. Both factors were used to stratify randomization of students.

The subjects spent a total of two hours in the experiment. After a brief introduction explaining the purpose of the experiment, they were given one-half hour to complete the ten-question test. At the completion of the test, they designated the five questions for which they had the least certainty about their answer. After a short break, they were oriented for 15 minutes to their computer and IR system, SWORD or BOOLEAN. Students then had up to 30 minutes to search for answers to the five questions for which they had greatest uncertainty about their original answers. They were required not only to answer each question, but also to give one or more document references that supported their answer.

The searching logs captured data about each query, including number of searches, total documents retrieved and viewed, and time taken. A *query* was defined as all of the interactions in attempting to find the answer to a question. A *search* was the entering of a search statement and retrieval of matching document titles. A document was considered *retrieved* if its title was in the list of document titles displayed after a search. A document was considered *viewed* if the user displayed the full text on the screen. For each user's query, we determined the number of searches, number of documents retrieved, and number of documents viewed. In addition to total number of searches, retrieved documents, and viewed documents for each query, we also calculated the number of each of these parameters required to reach an answer document.

The tests were scored independently by two members of the study team (WRH and SLW), whose interobserver agreement was good (kappa = 0.71). To assess information acquisition, a pre-test/post-test analysis was used. A McNemar's Test was performed for each test question, using data from those subjects who answered that question on the post-test.

RESULTS

A total of 13 subjects participated, six of whom used BOOLEAN and seven of whom used SWORD. There were no significant differences between the BOOLEAN and SWORD groups in computer experience or class rank. The average number correct on the initial ten-question test was 1.2, with no statistically significant difference between groups. The average number correct for the five questions searched upon was 4.1, again with no significant differences between groups (Table 2). Because there were no differences in general user characteristics or answers between the programs, the data were then pooled to determine information acquisition. Four of the ten questions showed a statistically significant difference in information found when using a searching program, while four others had a trend towards significance (Table 3).

Table 4 compares all of the questions in terms of searches done, documents retrieved, and documents viewed for each question, both in total as well as number required to retrieve an answer document. The majority of answer documents were found on the first search, within the top ten documents retrieved, and on the first document viewed.

Table 1: Ten questions for searching - answers in *italics*

1. A 60-year-old man from a poor socioeconomic environment is admitted with an acute illness characterized by mental disturbances, a sixth nerve palsy, and ataxia of gait. What specific emergency treatment is needed? *Thiamine.*
2. What percent of patients with Type II diabetes respond to oral hypoglycemic agents as their initial drug treatment? *60-70%.*
3. Mr. Rogers is seen in the Bend, OR Emergency Room. He states that he was bitten by a 'spider.' He is relatively certain that it was a black widow. What are the expected initial symptoms of the bite? *Muscular pain and rigidity.*
4. What organism is most commonly found in anaerobic osteomyelitis? *Bacteroides.*
5. You are seeing a diabetic man with severe gastropariesis. He has not improved on oral metoclopramide (Reglan) and was sent to you for additional treatment. What would you recommend? *Suppository form of metoclopramide.*
6. What electrocardiographic feature distinguishes Prinzmetal's angina from more typical angina pectoris? *ST elevation.*
7. Mrs. Towel, an 80-year-old woman on no medication, is seen for light-headedness and found to have a heart rate of 36 and third degree heart block. What is the most likely etiology of her heart block? *Lenegre's Disease* or *age-related changes in A-V conduction system.*
8. A strongly positive antibody test to which antigen is most typical of Mixed Connective Tissue Disease? *Anti-RNP antibody.*
9. What is the most common cause of sudden death among young athletes? *Hypertrophic cardiomyopathy.*
10. How is the organism which causes Rocky Mountain Spotted Fever transmitted? *Tick bite.*

Table 2: Test results for the study groups

	BOOLEAN	SWORD	Both
Number	6	7	13
Pre-Test Score (correct of 10)	1.8	1.6	1.7
Post-Test Score (correct of 5)	4.2	3.9	4.0

Table 3: Pre-Test/Post-Test results for each query

	Pre-Test		Post-Test		
Question	No. responses	% correct	No. responses	% correct	p
1	13	30.8	3	100	.08
2	13	23.1	6	83.3	.08
3	13	0	8	100	.005
4	13	23.1	9	100	.01
5	13	0	8	87.5	.008
6	13	0	12	100	.0005
7	13	0	4	25	.3
8	13	0	11	27.3	.08
9	13	15.4	1	100	.3
10	13	76.9	3	100	.08

We also performed a failure analysis of questions where the wrong answer was obtained, or where there was an unsuccessful retrieval or viewing (Table 5). Only four of the ten questions had any incorrect answers at all. The majority of these came from question 8, although almost all of those who got this question wrong retrieved the answer document, and over half viewed that document, indicating that perhaps it was a poorly worded question.

DISCUSSION

The purpose of this pilot study was to explore alternative methods of evaluating the performance of IR systems, based on ability to acquire information. Our results indicate that this approach is a viable alternative approach to measuring recall and precision, and may even be preferable, in that it indicates whether the searcher was able to use the system to find answers to questions. While this approach might not generalize to all uses of IR systems (i.e., the researcher who needed to find every relevant document on a topic), it appears to be appropriate for the specific questions that arise in the course of clinical practice [14].

One limitation of the study that was allowing subjects to choose only five questions to search. Not only did this make the statistical analysis more difficult, but it also made assessment of the adequacy of some questions difficult. In our next study, we will have users search on all questions in order to better assess the value of all questions searched by the IR system.

The next question is whether this approach will be able to allow comparison of different IR systems. To this end, we plan to compare two commercial MEDLINE systems that are used in the OHSU library, one of which features Boolean searching (*CD Plus*, CD Plus, Inc., New York, NY) and the other natural language searching (*Knowledge Finder*, Aries Systems, Inc., North Andover, MA) based upon clinical questions that were actually generated in the course of patient care during an information needs assessment study [15]. In this study, we will also compare these results with conventional recall-precision analysis.

ACKNOWLEDGEMENTS

This work was supported by Grant LM 05307 from the National Library of Medicine. The authors also thank Scientific American (New York, NY) for providing the text of *Scientific American Medicine* for this study.

REFERENCES

1. Marshall J. The continuation of end-user online searching by health professionals: Preliminary survey results. Abstracts of the Medical Library Association Annual Meeting. Detroit: Medical Library Association; 1990. Available from Medical Library Association.

2. Haynes RB, Ramsden MF, McKibbon KA, et al.: Online access to MEDLINE in clinical settings: Impact of user fees. Bulletin of the Medical Library Association. 1991; 79: 377-381.

3. Haynes RB, McKibbon KA, Walker CJ, et al.: Online access to MEDLINE in clinical settings. Annals of Internal Medicine. 1990; 112: 78-84.

4. Hersh WR, Hickam DH: A performance and failure analysis of SAPHIRE with a MEDLINE test collection. Journal of the American Medical Informatics Association. 1994; 1: 51-60.

5. Hersh WR, Hickam DH: A comparison of two methods for indexing and retrieval from a full-text medical database. Medical Decision Making. 1993; 13: 220-226.

6. Eisenberg M, Barry C: Order effects: a study of the possible influence of presentation order on user judgments of document relevance. Journal of the American Society for Information Science. 1988; 39: 293-300.

7. Schamber L, Eisenberg MB, Nilan MS: A re-examination of relevance: Toward a dynamic, situational definition. Information Processing and Management. 1990; 26: 755-776.

8. Hersh WR: Relevance and retrieval evaluation: perspectives from medicine. Journal of the American Society for Information Science. 1994; 45: 201-206.

9. Egan DE, Remde JR, Gomez LM, et al.: Formative design-evaluation of Superbook. ACM Transactions on Information Systems. 1989; 7: 30-57.

10. Mynatt BT, Leventhal LM, Instone K, et al.: Hypertext or book: which is better for answering questions? Proceedings of Computer-Human Interface 92. 1992: 19-25.

11. de Bliek R, Friedman CP, Wildemuth BM, et al.: Database access and problem solving in the basic sciences. Proceedings of the 17th Annual Symposium on Computers in Medical Care. 1993: 678-682.

12. Hersh WR, Hickam DH, Leone TJ: Word, concepts, or both: Optimal indexing units for automated information retrieval. Proceedings of the 16th Annual Symposium on Computers in Medical Care. 1992: 644-648.

13. Rubenstein R, Federman DD. "Scientific American Medicine." 1990 Scientific American. New York.

14. Gorman PN, Ash J, Helfand M, Beck JR: Assessment of information needs of primary care physicians. Proceedings of the Third Annual American Medical Informatics Association Spring Congress. 1992: 26.

15. Gorman P: Does the medical literature contain the evidence to answer the questions of primary care physicians? Preliminary findings of a study. Proceedings of the 17th Annual Symposium on Computers in Medical Care. 1993: 571-575.

Table 4: Searching results for all queries with both programs

Total searches done	
1	48
>1	17
Searches to find answer	
1st	51
After 1st	5
Not found	9
Total documents retrieved	
<=10	46
>10	19
Documents retrieved to find answer	
<=10	49
>10	7
Not found	9
Total documents viewed	
<=10	60
>10	5
Documents viewed to find answer*	
1	41
2-5	13
>6	5
Not found	6
Time per query (min.)	5.40

* There were three queries with answer documents viewed but not retrieved by searching due to answers being found by browsing through the database.

Table 5: Failure analysis

Query	Incorrect	Retrieved		Viewed	
		Yes	No	Yes	No
2	1	1	0	1	0
5	1	1	0	1	0
7	3	0	3	0	3
8	8	7	1	5	3
Total	13	9	4	7	6